International Journal of Applied Linguistics ◆ Vol. 18 ◆ No. 1 ◆ 2008

Investigating grammatical difficulty in second language learning: Implications for second language acquisition research and language testing

Rod Ellis *Chang Jiang Scholar of Shanghai International Studies University and University of Auckland*

Both second language acquisition (SLA) researchers and language testers collect data in order to make statements about what learners have learned. Many researchers and testers consider the ideal data for this purpose to be naturally occurring language use. This paper examines whether data elicited by instruments designed to provide separate measures of implicit and explicit second language knowledge afford a valid basis for determining what learners have learned. It reports on a study that tested predictions derived from Pienemann's Processability Theory regarding the learning difficulty of four grammatical structures. The results showed that the predictions were borne out in the data from the tests of implicit knowledge but not in the data from the tests of explicit knowledge. The study suggests that experimentally elicited data can be used to examine interlanguage development (i.e. how learners' implicit knowledge develops) and to make statements about learners' grammatical proficiency. It also indicates that what constitutes learning difficulty needs to be considered separately for implicit and explicit knowledge. The implications for SLA research and language testing are considered.[1]

Keywords: implicit knowledge, explicit knowledge, interlanguage development, SLA, language testing

第二语言习得的研究者和语言测试者收集数据是为了证明学习者学到的知识。为此目的，许多研究者和测试者认为最理想的数据应该来自于学习者在自然状态下使用的语言。本篇文章调查了用测试第二语言的隐性知识和显性知识的不同方法测出的数据是否为测试学习者学到的知识提供了有效的基础。本文阐述的研究成果是基于 Pienemann 提出的“语言可加工理论”检验四个语法结构学习难度的假设。本研究结果证实此项假设适用检测隐性知识而非显性知识。研究表明，通过实验法收集的数据可用于检测中介语的发展（例如，学习者的隐性知识如何得到发展）并且检验学习者的语法水平。此研究结果揭示，学习难度的构成应与隐性知识和显性知识区别对待。本文也提及了此项成果对第二语言习得研究和语言测试的意义。

*关键词：*隐性知识，显性知识，中介语发展，第二语言习得，语言测试

Introduction

What kinds of data are needed to determine what second language (L2) learners have learned? This is a question of importance for both second language acquisition (SLA) researchers and language testers. Two major kinds of data can be distinguished; (1) naturally occurring data (i.e. samples of language use in a real-life situation where the purpose was to satisfy some communicative and/or aesthetic need) and (2) elicited data (i.e. data collected by means of some specially designed instrument). Corder (1976) distinguished two types of elicited data: clinically elicited data, which involves 'getting the informant to produce data of any sort', and experimentally elicited data, where the aim is to get the informant to 'produce data incorporating particular features the linguist is interested in at the moment' (p. 69). Both SLA researchers and language testers have shown a general preference for naturally occurring data, on the grounds that these provide the most convincing evidence of what a learner knows and can do with the L2. The problem with such data, however, is that they do not readily afford the evidence needed to make statements about what learners have learned, especially if these statements concern the acquisition of specific grammatical structures (e.g. naturally occurring data is unlikely to contain plentiful examples of structures like hypothetical conditionals or oblique relative clauses). For this reason, there is a need to turn to elicited data, especially experimentally elicited data. So doing, however, raises another issue – the extent to which such data afford valid evidence of what learners have learned.

To address this issue it is necessary to refer to some theory of what is meant by 'learned'. The theory that I have proposed in a number of previous publications (e.g. Ellis 1993; 1994; 2004) is based on the distinction between implicit and explicit L2 knowledge. Theories of L2 acquisition (and, I would argue, current approaches to language testing) concur that it is implicit knowledge that is primary, as this is the type of knowledge that is involved in interlanguage development and that enables learners to engage in fluent communication. However, there is less agreement regarding the kind of data needed to examine learners' implicit knowledge. Generative theories of L2 acquisition (White 2003) have typically favoured a special type of experimentally elicited data (i.e. grammaticality judgements) whereas interactionist theories (Long 1996) have favoured naturally occurring or clinically elicited data.

The purpose of this paper is to examine to what extent experimentally elicited data can provide valid measures of learners' implicit knowledge. I shall attempt this by testing predictions from Pienemann's (1998; 2005) Processability Theory regarding the learning difficulty of a number of grammatical structures. Pienemann's theory has been previously tested using data collected from naturally occurring language use. Indeed, Pienemann has argued that this is the only kind of data that is appropriate for testing the

Table 1. Key characteristics of implicit and explicit knowledge

Characteristics	Implicit knowledge	Explicit knowledge
Awareness	Learner is intuitively aware of linguistic norms	Learner is consciously aware of linguistic norms
Type of knowledge	Learner has procedural knowledge of rules and fragments	Learner has declarative knowledge of grammatical rules and fragments
Systematicity	Knowledge is variable but systematic	Knowledge is often anomalous and inconsistent.
Accessibility	Knowledge is accessible by means of automatic processing	Knowledge is accessible only through controlled processing
Use of L2 knowledge	Knowledge is typically accessed when learner is performing fluently	Knowledge is typically accessed when learner experiences a planning difficulty
Self-report	Non-verbalizable	Verbalizable
Learnability	Potentially only learnable within the 'critical period'	Learnable at any age

theory. I will investigate whether elicited data obtained from an instrument designed to measure learners' implicit grammatical knowledge yields the same order of difficulty that Pienemann and other researchers have found in naturally occurring data. I will also investigate whether elicited data obtained from instruments designed to measure learners' explicit grammatical knowledge manifests the same or a different order of difficulty.

I will first provide definitions of implicit and explicit L2 knowledge, together with a brief account of Processability Theory. There follows a description of the instruments used to measure the two kinds of knowledge, together with a summary of previous studies which have demonstrated the validity of the measures obtained. The predictions of Processability Theory regarding the acquisition of four grammatical structures are then tested using the measures of implicit and explicit knowledge. Finally, the implications of the results for both SLA research and language testing are considered.

Defining implicit and explicit L2 knowledge

I have identified seven ways in which implicit and explicit knowledge of language can be distinguished (see Ellis 2004). These are summarized in Table 1. In line with these distinguishing characteristics the following definitions can be formulated:

> *Implicit knowledge* is intuitive, procedural, systematically variable, automatic and thus available for use in fluent, unplanned language use. It is not

verbalizable. According to some theorists it is only learnable before learners reach a critical age (e.g. puberty).

Explicit knowledge is conscious, declarative, anomalous and inconsistent (i.e. it takes the form of 'fuzzy' rules inconsistently applied), and is only accessible through controlled processing in planned language use. It is verbalizable, in which case it entails semi-technical or technical metalanguage. Like any type of factual knowledge, it is potentially learnable at any age.

These two types of knowledge are closely linked to implicit/explicit memory.[2] Eysenck (2003) defined *implicit memory* as 'memory that does not depend on conscious recollection' (p. 334) and *explicit memory* as memory that does. They are also closely related to Labov's (1970) sociolinguistic distinction between a *vernacular style* and *careful style* as they underlie the variability observed in learner performance.

The implicit/explicit distinction is not without controversy. One issue of debate is whether it constitutes a continuum or a dichotomy. Dienes and Perner (1999) claim that the distinction represents a continuum rather than a dichotomy, a position they see supported by Karmiloff-Smith's (1979) account of how implicit linguistic knowledge becomes progressively more explicit in children. However, others (e.g. Krashen 1981 and Schwartz 1993) have argued strongly that it constitutes a dichotomy. This view is supported by Paradis (1994; 2004), who presented evidence to suggest that the two types of knowledge are neurolinguistically distinct.

A further issue concerns the extent to which the distinction can be operationalized and thus investigated empirically. This is a matter of special importance for both SLA researchers and language testers. SLA researchers interested in instructed L2 acquisition, for example, need to determine whether instruction can only assist the development of explicit knowledge, as claimed by Krashen (1982), or whether it is able to contribute to implicit knowledge (see Doughty 2004). Testers need to be sure that what they are measuring is of relevance to 'proficiency' (i.e. the ability to use language in real-world contexts), and thus also need to establish whether their measures tap implicit knowledge (which is primary for many types of language use) or explicit knowledge.

DeKeyser (2003) has suggested that it might not be possible to distinguish the two types of knowledge empirically, given that explicit knowledge may be proceduralized to a degree that makes it functionally indistinguishable from implicit knowledge (i.e. both types of knowledge may be available in unplanned language use). The position I have adopted, however, is that advanced by Hulstijn (2002), namely that although it may be possible to 'speed up the execution of algorithmic rules to some extent' (p. 211), it is still possible to distinguish implicit and explicit knowledge.

A study by Han and Ellis (1998) lends some support to this position.They analysed scores derived from a battery of tests (an oral production test, a

timed grammaticality judgement test (GJT) and an untimed grammaticality test) and a measure of metalinguistic ability based on learners' verbalisations of a grammatical rule. In a Principal Component Factor Analysis, scores from the oral production test and the timed GJT loaded on one factor, while the untimed GJT and the metalinguistic comments score loaded on a second factor. Han and Ellis labelled these two factors 'implicit' and 'explicit L2 knowledge' respectively. This study was limited, however, in that it focused on a single grammatical structure (verb complementation). The study presented below is intended to further investigate the measurability of the two types of knowledge by examining a wider range of grammatical structures.

Processability Theory

Pienemann's Processability Theory seeks to explain what is known about acquisitional orders/sequences in terms of a set of processing procedures. As Pienemann (2005: 2) put it, 'once we can spell out the sequence in which language processing routines develop we can delineate those grammars that are processable at different points of development.' Drawing on Levelt's work on speech production, he proposed that language production, whether in the L1 or the L2, could only be explained with reference to a set of basic premises:

(1) speakers possess relatively specialized processing components that operate autonomously and in parallel;
(2) processing is incremental (i.e. a processor can start working on the incomplete output of another processor);
(3) in order to cope with non-linearity (i.e. the fact that the linguistic sequence does not match the natural order of events as in *Before the man rode off, he mounted his horse*), speakers need to store grammatical information in memory, and thus it follows that:
(4) grammatical processing must have access to a grammatical memory store, which Pienemann saw as task-specific and as involving 'procedural' rather than 'declarative' memory.[3]

It should be clear from this account that Processability Theory is in actuality a theory of language production. However, it can lay claim to being a theory of language acquisition in that it proposes that the processing procedures are hierarchical and are mastered one at a time. As Pienemann (2005: 13) put it, 'it is hypothesised that processing devices will be acquired in their sequence of activation in the production process.' Thus, the failure to master a low-level procedure blocks access to higher-level procedures and makes it impossible for the learner to acquire those grammatical features that depend on them.

Pienemann (1998; 2005) identified the following language generation processes:

1. word/lemma;
2. category procedure (lexical category);
3. phrasal procedures (head);
4. S-procedure and word order rule;
5. matrix/subordinate clause.

What distinguishes these processes is the nature of the grammatical information that the learner needs to deposit and exchange in what Pienemann calls 'feature unification'.

Initially, learners are unable to control any of the processes involved. At this stage learners are able to access L2 words, but these are invariant in form and are used in single-constituent utterances. The learners' lexicon is not annotated, while transfer of L1 annotation is blocked because the learner has not yet developed the specialized procedures to hold L2 grammatical information. Thus, the beginning learner 'is unable to produce any structures which rely on the exchange of specific grammatical information using syntactic procedures' (Pienemann 2005: 11).

The first procedure to be mastered is the 'category procedure'. Lexical entries are now annotated with a number of diacritic features (e.g. 'possessive' and 'number'). These can be accessed, but only within a single constituent, and are matched directly with the underlying conceptual content of a message, so no exchange of grammatical information is required. At this stage the learner is still not able to handle structures where diacritic features need to be matched across elements in a constituent or between constituents.

The ability to handle this begins at the next stage – the level of phrasal constituents. Thus, it is now possible for learners to handle such structures as articles, plural agreement (e.g. *many children*), and *do*-fronting (e.g. *Do he like it?*). Exchange of information in the phrasal procedure is required to check the value of a diacritic feature of one lexical entry (e.g. *child* – plural) with that of another (e.g. *many* – plural) to ascertain that they match and thus enable the production of a structural phrase (e.g. *many children*).

At this stage, however, exchange of information *between* structural phrases is still not possible. This is activated at the next stage – the S-procedure. This involves exchange of information between heads of different phrases, as in subject–verb agreement, which entails the unification of features such as person and number *across* constituent boundaries. The features of one constituent (the subject noun phrase) are deposited in the S-procedure and subsequently placed in another constituent (the verb phrase). When this becomes possible, learners are able to mark the 3rd person of the present simple tense with the *-s* morpheme.

The final procedure to be acquired enables learners to process the word order of subordinate structures such as that found in embedded questions in English (e.g. *He asked where I lived*) and verb-end in German (e.g. *Er fragt warum ich traurig war*).

Pienemann argues there is a basic difference between the first three procedures and the last two in the hierarchy, in that structures appearing in levels 1–3 cannot be represented by constituent structure rules because the S-procedure has not been developed. Thus, in the early stages 'sentences are formed using simplified procedures based on a direct mapping of argument structure onto functional structure' (2005: 14). According to this theory, then, learning difficulty and the sequence of acquisition are determined by the nature of the processing procedure required to produce a specific grammatical feature.

Measuring implicit and explicit L2 knowledge

In Ellis (2005) I reported a study that sought to identify measures of implicit and explicit knowledge that were relatively independent of each other. Five tests were investigated. Table 2 provides a summary description of these five tests. Each test was designed to provide a measure of learners' knowledge of 17 grammatical structures, selected to reflect different levels of learning difficulty. Three of the tests (the Oral Imitation Test, the Oral Narrative Test and the Timed Grammaticality Judgement Test) were designed to provide measures of the learners' L2 implicit knowledge, while the other two tests (the Untimed Grammaticality Judgement Test and the Metalinguistic Test) were designed to provide measures of L2 explicit knowledge. The design of the tests was based on criteria derived from the characteristics of the two types of knowledge described in Table 1.

In Ellis (2005), data collected from 111 mixed proficiency learners were subjected to a Principal Component Factor Analysis. In a two-factor solution, the Oral Imitation Test, The Oral Narrative Test and the Timed GJT loaded on one factor, while the Untimed GJT and the Metalinguistic Test loaded an the other factor. No test loaded strongly on both factors. This result suggested that the tests were largely measuring different constructs in accordance with their design, and indicated that it was possible to obtain relatively separate measures of implicit and explicit knowledge. In Ellis and Loewen (2007), the same data were submitted to a Confirmatory Factor Analysis. This showed that a solution based on the implicit/explicit distinction (as in the earlier study) was statistically significant and more satisfactory than a solution based on the distinction between 'production' (i.e. the Oral Imitation Test and the Oral narrative Test) and 'decision' (i.e. the Timed GJT, the Untimed GJT and the Metalinguistic Test). In other words, the best available interpretation of the data was one that accorded with the theoretical basis of the design of the testing instruments – the distinction between implicit and explicit knowledge.

In Ellis (2006), a larger sample of 224 learners completed four of these tests (the Oral Narrative Test was not included[5]). The sample was made up of a number of different groups. The majority (N = 147) were international

Table 2. Description of five grammar tests

Test	Description	Reliability (Cronbach alpha)
Oral Imitation Test	A total of 34 sentences (two per structure, one grammatical and one ungrammatical). Test-takers listened to each sentence and first indicated whether they agreed or disagreed with the proposition it expressed. They then attempted to repeat the sentence correctly and were audio recorded. Scoring was based on whether learners successfully repeated/corrected the target structure in each sentence. A percentage accuracy score was calculated.[4]	.88
Oral Narrative Test	A story designed to provide obligatory occasions for the use of a number of grammatical structures (e.g. regular past tense and 3rd person *-s*) was read twice by the participants, who were then asked to retell the story orally within 3 minutes. Obligatory occasion analysis was then used to establish the accuracy with which the participants used each structure.	.85 (interrater agreement)
Timed Grammaticality Judgement Test	This was a computer-delivered test consisting of 68 sentences (4 per structure, 2 grammatical and 2 ungrammatical). Test-takers were required to indicate whether each sentence was grammatical or ungrammatical by pressing response buttons within a fixed time limit. The time limit for each sentence was based on piloting with native speakers. Each item was scored dichotomously as correct/incorrect (reflecting the responses of native speakers), with items not responded to scored as incorrect. A percentage accuracy score was calculated.	.81
Untimed Grammaticality Judgement Test	Same computer-delivered content as the Timed GJT. Test-takers were required to indicate in their own time whether each sentence was grammatical or ungrammatical. This test provided a percentage judgement accuracy score based on the participants' dichotomous responses. Total accuracy scores as well as separate scores for the grammatical and ungrammatical sentences were calculated.	.83

Table 2 *Continued*

Test	Description	Reliability (Cronbach alpha)
Metalinguistic Knowledge Test	This test consisted of two parts (but only part 1 is described as only the scores from this part were used in this study). This presented test-takers with 17 ungrammatical sentences (one sentence per structure), and required them to select the rule that best explained each error out of four choices provided, as in this example: *You must to wash your hands before eating.* a. *Must to* is the wrong form of the imperative b. Change to *must have to wash* to express obligation. c. Modal verbs should never be followed by a preposition. d. After *must* use the base form of the verb not the infinitive. A total percentage accuracy score was calculated.	.90

students of mixed language proficiency, mainly from China, who were studying English as a second language either in a language school in New Zealand or as part of an undergraduate degree programme at the University of Auckland. A small group (N = 28) were first-year Japanese students at an all-women's university in Tokyo. With a few exceptions, these students had very limited procedural ability in English. A third group (N = 54) were students enrolled in a four-year BEd TESOL programme in Malaysia. These students had undergone an intensive English preparation course, and generally spoke and wrote English fluently and with confidence.

Overall, then, the English proficiency of the learners in this sample was very mixed, ranging from false beginners to advanced learners displaying high levels of linguistic competence and fluency. A Principal Components Factor Analysis (stipulating a two-factor solution) was carried out on this sample (with missing cases omitted). The results are shown in Table 3 below. The Untimed GJT (ungrammatical sentences) and Metalinguistic Test loaded on factor 1 while the Oral Imitation Test and the Timed GJT (all sentences) loaded on factor 2, with no cross-loadings of any significance. Thus, for this larger sample, the analysis again supported the claim that tests were measuring two relatively distinct constructs – explicit and implicit knowledge.

Table 3. Structure matrix for the Principal Component Analysis (oblique with Kaiser normalization)

Component	Total	% of variance	Cumulative %
1	2.426	60.64	60.64
2	.801	20.03	80.67
Test		**Component 1**	**Component 2**
Imitation Test		.121	.827
Timed GJT		.079	.956
Untimed GJT (ungram.)		.876	−.025
Metalinguistic knowledge		.897	.014

Using experimentally elicited data to test predictions about learning difficulty

Processability Theory affords a basis for predicting which structures the learners in Ellis's (2006) study found easy and which difficult. Four structures from the 17 structures for which data were available from the battery of tests described above were chosen to represent each of the hierarchical processing operations distinguished by the theory (see earlier description of these). Information about these four structures can be found in Table 4.

For the category procedure, possessive *-s* was chosen. Pienemann (2005) claims that possessive *-s* is a feature marked diacritically in lexical entries, and thus can be accessed directly from the learner's lexicon. For the phrasal procedure, *since/for* was chosen. Here the selection of preposition depends on the nature of the following noun phrase. If this refers to a specific point in time, *since* is required (e.g. *since 1985*), while if it refers to a period of time *for* is required (e.g. *for five years*). Thus, choice is entirely dependent on information contained within the phrasal constituent. For the S-procedure, 3rd person *-s* was chosen. Pienemann (2005) gives this as an example of this procedure. Finally, for subordinate clause procedure, question tags were chosen; the form of a question tag depends on information contained in the main clause that precedes it.

The participants were the same as in Ellis (2006: see previous section). For one of the analyses reported below (involving implicational scaling), however, 20 participants were randomly selected from the total sample.

Research based on Processability Theory has used 'emergence' as the measure of acquisition. That is, researchers have considered a feature 'acquired' if a learner has used it in two non-formulaic utterances. As noted previously, it is also clear that the theory addresses acquisition in relation to learner production. For this reason, I elected to use only the Oral Imitation

Table 4. The four grammatical structures

	Description	Typical learner error
Possessive *-s*	*-s* is attached to a modifying noun to signal it is the possessor.	**Liao is still living in his rich uncle house.*
Question tags	The choice of auxiliary in a question tag is dependent on the form of the main verb (e.g. if the main verb contains an auxiliary then the same auxiliary must be chosen in the question tag).	**We will leave tomorrow, isn't it?*
Since/for	*Since* denotes a period of time commencing at a specific point in the past and continuing into the present; *for* is used when the period is denoted in terms of a number of time units.	**He has been living in New Zealand since three years.*
3rd person *-s*	*-s* is attached to the base form of the verb in the 3rd person of the Present Simple Tense.	**Hiroshi live with his friend Koji.*

Table 5. Means implicit and explicit scores for four structures

Processing procedure	Structure	Mean implicit score	Mean explicit score
Subordinate clause procedure	Question tags	.41 (4)	.75 (2)
-s procedure	3rd person *-s*	.46 (3)	.64 (4)
Phrasal procedure	*Since/for*	.52 (2)	.72 (3)
Category procedure	Possessive *-s*	.61 (1)	.81 (1)

Test scores as the measure of implicit knowledge. I used the scores for the ungrammatical sentences of the Untimed Grammaticality Judgment Test as the measure of explicit knowledge, as the previous studies had shown this to be the best measure of this type of knowledge. Two analyses are reported. The first is based on mean accuracy scores. The second used implicational scaling. In both cases I expected to find that the theory successfully predicted learning difficulty as implicit knowledge but not necessarily as explicit knowledge.

Table 5 shows the mean implicit and explicit scores for each of the four structures for the whole sample. The rank order of difficulty of the four structures, as shown in the implicit scores, conforms to the order of difficulty

predicted by Processability Theory. That is, possessive *-s* emerges as the easiest structure, *since/for* as next, 3rd person *-s* as next and question tags as the most difficult. However, a very different order of difficulty is evident for the explicit scores. While possessive *-s* was still the easiest (not surprisingly, perhaps, as the rule for this structure is conceptually simple and requires little metalanguage), question tags proved the next easiest, with 3rd person *-s* and *since/for* more difficult.

Mean scores, however, can be misleading. It does not follow that individual learners will experience the same order of difficulty as the sample as a whole. It is for this reason that research based on the Processability Theory has invariably examined the sequence of acquisition of individual learners. Accordingly, a second analysis using implicational scaling was undertaken. A structure was considered acquired as implicit knowledge if a learner used it correctly in both the items measuring it in the Oral Imitation Test. This scoring decision reflected Pienemann's use of 'emergence' as a criterion of acquisition. A structure was considered acquired as explicit knowledge only if the learner correctly judged both of the two ungrammatical sentences measuring it.[6] The results of the implication scaling are shown in Table 6.

Table 6. Implicational scaling of 20 learners' implicit and explicit knowledge

Learner	Q tag		3 per. *-s*		*Since/for*		Poss. *-s*	
	Impl.	Expl.	Impl.	Expl.	Impl.	Expl.	Impl.	Expl.
30	–	+–		+	–	–	–	–
71	–	+–		+	+	+	–	–
121	–	+	–	–	+	–	–	+
133	–	+	–	–	–	–	–	–
187	–	–	–	–	–	–	–	–
98	–	+	+	+	–	+	+	–
179	–	+–		+	–	+	+	–
5	–	+–		+	+	+	+	+
18	–	+	–	–	+	+	+	+
42	–	+–		+	+	–	+	+
143	–	+–		+	+	+	+	–
171	+	+	–	+	+	+	+	+
10	–	+	+	–	+	–	+	+
16	–	+	+	+	+	+	+	–
27	–	+	+	–	+	–	+	+
39	–	+	+	+	+	–	+	+
216	–	+	+	+	+	+	+	+
169	+	+	+	+	+	+	+	+
198	+	+	+	+	+	+	+	+
208	+	+	+	+	+	+	+	+
Total	4	19	9	15	15	12	15	12

The coefficient of scalability for the measure of implicit knowledge was .80, reaching the criterion level recommended by Hatch and Farhady (1982). The four structures scaled as predicted for the Oral Imitation Test scores. That is, with very few exceptions, learners who have acquired question tags have also acquired the other three structures, learners who have acquired 3rd person *-s* have also acquired the other two structures, and learners who have acquired *since/for* have also acquired possessive *-s*. The reverse is clearly not true. Acquisition of possessive *-s*, for example, does not implicate acquisition of the other structures. In other words, the implicational scaling of the four structures lends strong support to the claims about learning difficulty derived from Processability Theory. Thus, the data obtained from the Oral Imitation Test proved comparable to the unplanned language use data that Processability researchers have traditionally collected. That is, they allowed predictions based on the theory to be successfully tested.

In contrast, as expected, Processability Theory does not predict learning difficulty as explicit knowledge. Indeed, the structure predicted to constitute the greatest difficulty in learning (question tags) emerged as the simplest of the four structures in the Untimed Grammaticality Judgment Test, with 19 learners judging the ungrammatical sentences correctly.

The difference between the measures of implicit and explicit knowledge is also evident in other ways. For example, whereas 15 of the 20 learners have acquired the two simplest structures (possessive *-s* and *since/for*) as implicit knowledge, only 12 have acquired these structures as explicit knowledge. In contrast, whereas only 4 and 9 learners respectively have acquired the two most difficult structures as implicit knowledge, 19 and 15 learners have acquired these structures as explicit knowledge. One learner (No. 187) has neither implicit nor explicit knowledge of any of the four structures. Three learners (Nos. 169, 198, 208) have both implicit and explicit knowledge of all four structures. Four learners (Nos. 30, 71, 121, 133) have no implicit knowledge but explicit knowledge of at least one of the structures. Many learners (e.g. Nos. 5 and 216) have explicit knowledge of a structure without implicit knowledge, and somewhat fewer have implicit knowledge without any explicit knowledge of a structure.

Conclusion

The analysis presented above demonstrates that learning difficulty (as defined by Processability Theory) varied according to the type of knowledge being measured, reinforcing the conclusion drawn by Ellis's (2006) study of the learning difficulty of all 17 structures investigated. The study investigated learning difficulty in relation to the grammatical structures of English, but the findings are in principle relevant to other languages as well. Indeed, Pienemann's Processability Theory is a general theory of L2 acquisition, which has been tested on a number of different languages. In my final

comments, therefore, I will consider a number of general implications of the main finding of this study for both SLA researchers and for language testing.

SLA has been and still is primarily concerned with how learners develop linguistic competence, especially grammatical competence. While it is true that attention in the last decade or so has also been given to examining how pragmatic competence (see e.g. Kasper and Rose 2002) or conversational competence (Markee 2000) develops, the focus remains on how learners acquire linguistic resources. Indeed, pragmatic and conversational competence are realized primarily by means of linguistic resources, and thus (to my mind) it is inevitable that researchers will continue to focus on how these are acquired. Irrespective of the theoretical paradigm that informs SLA, linguistic competence is understood as consisting of implicit knowledge. This is what researchers are talking about when they refer to 'interlanguage'. Therefore, how best to obtain measurements of implicit knowledge remains a key issue, keenly debated, in SLA circles (see e.g. Norris and Ortega 2003 and Doughty 2003). The generally preferred means is to tap learners' unplanned communicative language use – what Norris and Ortega (2000) called 'free constructed response'. Surprisingly, however, few published SLA studies have been based on such data; many studies continue to measure acquisition by means of untimed metalinguistic judgments, selected responses or constrained constructed responses. One reason for this is the difficulty of obtaining adequate exemplars of the specific structures targeted for study from tasks designed to elicit free constructed responses. However, as the results of the analysis reported above suggest, tests such as the Untimed GJT that encourage learners to focus on formal accuracy rather than message conveyance permit learners to access their explicit knowledge, and cannot therefore convincingly shed light on their interlanguages. What is needed is an instrument that will gather information about learners' implicit knowledge of specific linguistic features. The kind of Oral Imitation Test described in Table 2 is promising in this respect. This test appears capable of producing data with the same essential characteristics as the free constructed response data which Processability research has utilized.

The Oral Imitation Test may also provide a much-needed means of determining the stage of development that individual learners have reached. The need for some general index of development has long been recognized (Larsen-Freeman 1978), but none of those suggested to date has been adopted. It might be possible to design an Oral Imitation Test to measure learners' implicit knowledge of grammatical features that have been carefully chosen to represent the different processing procedures that Pienemann has shown to characterize L2 development. If this does prove possible, it would enable SLA researchers to describe a learner's developmental stage in much more precise terms than the currently crude way it is described in so many studies (i.e. as 'beginner', 'intermediate' or 'advanced').

I am somewhat more hesitant to propose any applications of the tests I employed in the study reported above to language testing. There are two

reasons for this. One is simply that I am less familiar with the language testing literature than with the SLA literature. The other is that I am aware that, given current trends in language testing, language testers are unlikely to look favourably on the kinds of tests I have used in the study reported above. These trends favour either an approach to testing based on a model of communicative competence (e.g. Bachman and Palmer 1996) or on specific tasks replicating real-life activities (e.g. Douglas 2000; McNamara 1996). Both approaches emphasize what learners can *do* with language rather than what they *know*. In such tests, measurements are not derived directly from the tests themselves but indirectly from ratings of the performances elicited by the tests. One of the key motives for such tests is the positive 'backwash' effect they are likely to have on teaching. Clearly, none of the tests described in Table 2 reflects these approaches; they are directed at what learners know, not at what they can do with language; they do not correspond to the kinds of activity favoured in current language pedagogy, and thus have the potential for negative backwash.

There are, however, a number of ways in which language testers might benefit from the psycholinguistic approach to testing I have explored in this paper. First, I would argue that the distinction between implicit and explicit knowledge is as important for language testers as it is for SLA researchers. Yet, to the best of my knowledge, language testers have paid no attention to it. To what extent do the kinds of task popular in performance-based testing and the conditions under which they are performed tap the test-taker's implicit or explicit knowledge? If different types of language use (e.g. planned vs. unplanned) vary in terms of the utility of these two types of knowledge, do the assessment tasks used to measure them reflect this variation accurately? In other words, language testers need to address what type of linguistic knowledge their tests measure in order to establish construct validity. Are language testers solely interested in implicit knowledge? Is there a case for also obtaining measures of learners' explicit knowledge?

Secondly, language testers cannot ignore linguistic competence (and indeed have not done so). What learners can do with language is to a very considerable extent dependent on what language they know. However, there appears to exist a tension in language testing circles arising from the desire to measure functional ability while at the same time assessing test-takers' linguistic accuracy (McNamara 1996). This tension, it seems to me, derives from the difficulty of correlating descriptors of functional ability with linguistic exponents. It is clearly evident in the attempts to develop tests based on the Common European Framework of Reference for Languages. What, for example, are the linguistic correlates of this 'can do' statement for the B1 level of 'overall spoken interaction'?

> Can exploit a wide range of simple language to deal with most situations likely to arise whilst travelling. Can enter unprepared into conversation

> on familiar topics, express personal opinions and exchange information on topics that are familiar, of personal interest or pertinent to everyday life (e.g. family, hobbies, work, travel and current events) (Council of Europe: 74).

Such a description may enable a tester to identify what vocabulary should be tested, but it provides no clues whatsoever about what grammatical structures are involved. Nor does the Common European Framework document contain any clues. What is needed is a research project aimed at matching 'developmental levels' (as described in Processability Theory, for example) with 'proficiency levels'. But there have been few attempts to do so, reflecting perhaps the separateness of SLA and language testing as fields of study. One possibility might be to provide measures of both development (by means of a test such as the Oral Imitation Test) and proficiency (by means of performance-based testing). In other words, language testers might like to consider assessing both knowledge and functional ability, and might find that separate tests will work better – at least until our understanding of the correlation between them is more complete.[7]

However, the most obvious application of the results of the study reported in this paper lies in the testing of grammar. Purpura (2004) notes that language testers have neglected the assessment of grammatical ability in recent years, and that what grammar tests there have been have changed little since the 1960s – that is, they address morphosyntactic form by means of selected response or limited-production formats. Purpura argues that 'separate tests (or subtests) such as these provide only a partial measure of grammatical ability' and 'are perceived by current and past students, teachers, administrators and content teachers as being "old-fashioned" and "out of touch" with their learning goals' (p. 253). Although Purpura acknowledges a role for discrete-point testing, he clearly favours testing grammar by means of tasks that elicit more authentic language use. The problem with such 'focused tasks', however, is that they frequently fail to elicit the structure(s) they have targeted, as learners are adept at avoiding difficult structures, especially in a testing context (see Ellis 2003). What is really needed in grammar testing is a test-type that measures the kind of procedural ability tapped by tasks but also provides guaranteed information about learners' knowledge of specific target structures. The Oral Imitation Test, I would argue, achieves this. Furthermore, such a test, as already noted, can be designed to include grammatical structures representative of different levels of development. If a separate measure of the test-taker's explicit knowledge is required, then, a test requiring learners to judge (and perhaps correct) ungrammatical sentences containing the target structures would work well. Such tests, of course, are unlikely to be self-standing, but could figure in a battery of tests that should also include performance tasks.

Notes

1. This study was made possible by a Marsden Grant from the New Zealand Royal Society of Arts. A team of researchers contributed to the research reported in this article: Catherine Elder, Shawn Loewen, Rosemary Erlam, Ute Knoch, Jenefer Philp and Satomi Mizutani.
2. Studies such as Tulving and Schachter's (1990) suggested that implicit memory may consist of multiple systems rather than a single system. It is possible, therefore, that implicit linguistic knowledge is better conceptualized in terms of separate stores for a perceptual and a conceptual system.
3. Eysenck (2003) notes that the procedural/declarative memory and implicit/explicit memory cannot be clearly distinguished. That is, to all intents and purposes they should be considered as referring to the same mental phonemena.
4. Elicited imitation has been widely used in SLA research. Readers are referred to Bley-Vroman and Chaudron (1994) and Vintner (2002) for reviews of the literature relating to this method. In the Oral Imitation Test in this study, learners were asked to repeat the sentences they heard correctly. They were not warned that some of the sentences contained errors. If they repeated the error they were scored 0. To score 1 for a sentence they needed to perform the target structure correctly; they were not required to produce the whole sentence correctly. The length of the sentences was carefully controlled to make it difficult for the learners to simply memorize them. Also, because the sentences were presented as 'belief statements', to which they had to respond by indicating whether they agreed or disagreed before they attempted to repeat the sentences, the likelihood of their attempting to memorize the sentences was further reduced. A detailed rationale for the Oral Imitation Test is provided in Erlam (2006).
5. The test excluded from this study was the Oral Narrative Test. This test provided measures of only a subset of the 17 structures, and for this reason was not included.
6. In accordance with the results of the Principal Component Analysis shown in Table 3, only the ungrammatical sentences from the untimed GJT were used to derive explicit knowledge scores.
7. It is not clear to me, however, that it will ever prove possible to successfully correlate 'developmental' and 'functional' descriptors of language proficiency even if developmental features are taken to include non-target language forms (which language testers may not find acceptable). This is a case of apples and oranges.

References

Bachman, L. and A. Palmer (1996) *Language Testing in Practice: Designing and Developing Useful Language Tests*. Oxford: Oxford University Press.

Bley-Vroman, R. and C. Chaudron (1994) Elicited imitation as a measure of L2 competence. In E. Tarone, S. Gass and A. Cohen (eds)., *Research Methodology in SLA*. Hillsdale, NJ: Erlbaum. 245–61.

Council of Europe. Common European Framework of Reference for Languages, www.coe.int/T/E/Cultural_Co-operation/education/. Accessed February 2006.

Corder, S.P. (1976) The study of interlanguage. In *Proceedings of the Fourth International Conference of Applied Linguistics*. Munich: Hochschulverlag.

DeKeyser, R. (2003) Implicit and explicit learning. In C. Doughty and M. Long (eds.), *Handbook of Second Language Acquisition*. Malden, MA: Blackwell. 310–48.
Dienes, Z. and J. Perner (1999) A theory of implicit and explicit knowledge. *Behavioural and Brain Sciences* 22: 735–808.
Doughty, C. (2003) Instructed SLA: constraints, compensation and enhancement. In C. Doughty and M. Long (eds.), *The Handbook of Second Language Acquisition*. Malden, MA: Blackwell. 256–310.
— (2004) Effects of instruction on learning a second language: a critique of instructed SLA research. In B. Vanpatten, J. Williams and S. Rott (eds.), *Form–Meaning Connections in Second Language Acquisition*. Mahwah, NJ: Erlbaum. 181–202.
Douglas, D. (2000) *Assessing Languages for Specific Purposes*. Cambridge: Cambridge University Press.
Ellis, R. (1993) Second language acquisition and the structural syllabus. *TESOL Quarterly* 27: 91–113.
— (1994) A theory of instructed second language acquisition. In N. Ellis (ed.), *Implicit and Explicit Learning of Languages*. San Diego, CA: Academic Press. 79–114.
— (2003) *Task-Based Language Learning and Teaching*. Oxford: Oxford University Press.
— (2004) The definition and measurement of explicit knowledge. *Language Learning* 54: 227–75.
— (2005) Measuring implicit and explicit knowledge of a second language: a psychometric study. *Studies in Second Language Acquisition* 27: 141–72.
— (2006) Modelling learning difficulty and second language proficiency: the differential contributions of implicit and explicit Knowledge. *Applied Linguistics* 27: 431–63.
— and Loewen, S. (2007) Confirming the operational definitions of explicit and implicit knowledge in Ellis (2005): responding to Isemonger. *Studies in Second Language Acquisition* 29: 119–26.
Erlam, R. (2006) Elicited imitation as a measure of L2 implicit knowledge: an empirical validation study. *Applied Linguistics* 27: 464–91.
Eysenck, M. (2003) *Principles of Cognitive Psychology*. Hove, UK: Psychology Press.
Han, Y. and R. Ellis (1998) Implicit knowledge, explicit knowledge and general language proficiency. *Language Teaching Research* 2: 1–23.
Hatch, E. and H. Farhady (1982) *Research Design and Statistics for Applied Linguistics*. Rowley, MA: Newbury House.
Hulstijn, J. (2002) Towards a unified account of the representation, processing and acquisition of second language knowledge. *Second Language Research* 18: 193–223.
Karmiloff-Smith, A. (1979) Micro- and macro-developmental changes in language acquisition and other representation systems. *Cognitive Science* 3: 91–118.
Kasper, G. and K. Rose (2002) *Pragmatic Development in a Second Language*. Oxford: Blackwell.
Krashen, S. (1981) *Second Language Acquisition and Second Language Learning*. Oxford: Pergamon.
— (1982) *Principles and Practice in Second Language Acquisition*. Oxford: Pergamon.
Labov, W. (1970) The study of language in its social context. *Studium Generale* 23: 30–87.
Larsen-Freeman, D. (1978) An ESL index of development. *TESOL Quarterly* 12: 439–48.
Long, M. (1996) The role of the linguistic environment in second language acquisition. In W. Ritchie and T. Bhatia (eds.), *Handbook of Second Language Acquisition*. San Diego, CA: Academic Press. 413–68.
Markee, N. (2000) *Conversation Analysis*. Mahwah, NJ: Erlbaum.

McNamara, T. (1996) *Measuring Second Language Performance*. London: Longman.
Norris, J. and L. Ortega (2000) Effectiveness of L2 instruction: a research synthesis and quantitative meta-analysis. *Language Learning* 50: 417–528.
— (2003) Defining and measuring SLA. In C. Doughty and M. Long (eds.), *Handbook of Second Language Acquisition*. Malden, MA: Blackwell. 717–60.
Paradis, M. (1994) Neurolinguistic aspects of implicit and explicit memory: implications for bilingualism and second language acquisition. In N. Ellis (ed.), *Implicit and Explicit Language Learning*. London: Academic Press. 393–419.
— (2004) *A Neurolinguistic Theory of Bilingualism.* Amsterdam: Benjamins.
Pienemann, M. (1998) *Language Processing and Second Language Development: Processability Theory*. Amsterdam: Benjamins.
— (2005) An introduction to processability theory. In M. Pienemann (ed.), *Cross-linguistic Aspects of Processability Theory.* Amsterdam: Benjamins. 1–60.
Purpura, J. (2004) *Assessing Grammar*. Cambridge: Cambridge University Press.
Schwartz, B. (1993) On explicit and negative data effecting and affecting competence and linguistic behavior. *Studies in Second Language Acquisition* 15: 147–63.
Tulving, E. and D. Schachter (1990) Priming and human memory. *Science* 247: 301–6.
Vintner, T. (2002) Elicited imitation: a brief overview. *International Journal of Applied Linguistics* 12: 54–73.
White, L. (2003) *Second Language Acquisition and Universal Grammar*. Cambridge: Cambridge University Press.

e-mail: r.ellis@auckland.ac.nz *[Received November 1, 2007]*